Journal of Hydrology 535 (2016) 699–709

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Including land use information for the spatial estimation of groundwater quality parameters – 2. Interpolation methods, results, and comparison

C.P. Haslauer [a,*], T. Heißerer [b], A. Bárdossy [b]

[a] University of Tübingen, Center for Applied Geoscience, WESS, Hölderlinstr. 12, 72076 Tübingen, Germany
[b] University of Stuttgart, Institute for Modelling Hydraulic and Environmental Systems, Department of Hydrology and Geohydrology, Pfaffenwaldring 61, 70569 Stuttgart, Germany

ARTICLE INFO

Article history:
Received 29 June 2015
Received in revised form 8 January 2016
Accepted 19 January 2016
Available online 29 January 2016
This manuscript was handled by Peter K. Kitanidis, Editor-in-Chief, with the assistance of Roseanna M. Neupauer, Associate Editor

Keywords:
Geostatistics
Interpolation
Secondary information
Copula
Land use
Categorical variable

SUMMARY

Two dominant processes determine solute concentration in groundwater: vertical infiltration and horizontal advection. The goal of this paper is to incorporate both processes into a geostatistical model for spatial estimation of solute concentrations in groundwater. A multivariate copula-based methodology is demonstrated that considers infiltration via the marginal distribution and solute transport via the multivariate spatial dependence structure.

The novel approach is compared to traditional methods as Ordinary- and External Drift Kriging. Leave-one-out cross-validation demonstrates that the novel approach estimates better both in concentration and in probability space, and improves the quantification and quality of uncertainty. The gain in uncertainty reduction is equivalent to at least a few hundred additional observations when Ordinary Kriging was used.

Both censored and not-censored measurements are included. An ideal neighborhood size is estimated via cross-validation. The methodology is general and can incorporate other kinds of secondary information. It can be used to evaluate effects of land use changes.

1. Introduction

Estimating a variable's value at an un sampled location is the fundamental problem of geostatistics. Generally, the estimation is based on measurements in the vicinity. Additionally, readily available secondary information should be considered to enhance the estimation, as long as primary and secondary variables are related. For example, geodetic elevation can improve the estimate of precipitation intensities (Bárdossy and Pegram, 2013; Goovaerts, 2000) or geophysical measurements of electrical conductivity can improve the estimates of salt concentrations in groundwater (Pozdnyakova and Zhang, 1999; Shim et al., 2004).

In this paper, the composition of land use in the vicinity of an interpolation location (Heisserer et al., 2016) serves as secondary information to enhance the estimation of groundwater quality parameters. The spatial dependence is described in a geostatistical model using copulas (Bárdossy, 2011) that is extended to incorporate secondary information. The proposed methodology includes the two key processes responsible for anthropogenic contamination in shallow groundwater: a dominantly vertical process (infiltration) via locally mixed distributions that reflect the composition of land use in the vicinity and a dominantly horizontal process (lateral transport processes) via the geostatistical model of spatial dependence. Copulas offer the unique opportunity to couple both processes as they couple the marginal distributions and the multivariate distribution that describes here the spatial dependence. Groundwater flow advects solute concentrations laterally at relatively large scales (here ~40 km). Infiltration acts on much smaller scales (here ~1.5 km).

The following properties are desirable for theoretical geostatistical models (see also Zhu and Journel, 1993):

- The uncertainty of the model should be quantifiable on different scales.
- Constraints on the distribution (e.g. non-negative constraints, certain marginal distributions) should be possible.
- Hard constraints in the data (e.g. censored measurements) should be reflected in the model adequately.
- The model should be able to include secondary information or combinations of multiple types of secondary information (e.g. co-variables, soft prior information).

The methods presented in this paper advance all four attributes of geostatistical models. Basic geostatistical models incorporate "crisp" cardinal measurements on the ratio scale. This paper shows

* Corresponding author. Tel.: +49 (0)7071 29 73081.
E-mail address: claus.haslauer@uni-tuebingen.de (C.P. Haslauer).

http://dx.doi.org/10.1016/j.jhydrol.2016.01.054
0022-1694/

a methodology that allows various "other" or "secondary" information to be incorporated in geostatistical models. Such other information can be measured either on an interval scale or on a nominal scale. A typical example for measurements on an interval scale are so called "censored" measurements (in contrast to "crisp"), commonly referred to as "measurements below detection limit". For those measurements it is known that the true measurement values reside in an interval between zero and their analytical detection limit – this is valuable information that should not be discarded but utilized (Bárdossy, 2011). A typical example for measurements on the nominal scale (commonly referred to as "soft" information) is the type of land use at the measurement location. Generally, the incorporation of soft categorical information into geostatistical models is difficult, because it is not ordered and hence there is no covariance. Both secondary information and censored measurements are taken into account for an improved geostatistical model.

The basis of spatial interpolation using copulas has been laid by Bárdossy and Li (2008). Censored measurements can be included using the probability of non-exceedance (Bárdossy, 2011). Depending on the contaminant, different kinds of secondary information could improve the geostatistical model. In this paper, land use categories are used to improve the estimation of anthropogenic contaminants. Similarly, information on the spatial distribution of hydrogeological units could improve the interpolation of geogenic contaminants. A measure of information content that quantifies the worth of a certain type of secondary information for a certain contaminant is presented by Heisserer et al. (2016).

The copula-based methods presented are compared to traditional geostatistical interpolation methods such as Ordinary Kriging (OK, without secondary information) and External Drift Kriging (EDK, with secondary information). Those traditional methods rely on assumptions of stationarity that cannot be expected at the scale of the state of Baden-Württemberg, Germany: for example, the composition of land use varies between different areas of the state. In a predominantly forested area such as the Black Forest, lower nitrate concentrations can be expected compared to an area with mixed and agricultural land use such as the Kraichgau (Fig. 1). However, for water management purposes, there is the need to interpolate at the State scale. By incorporating the secondary land use information via the marginal distribution into copula-based geostatistical models, non-stationary effects can be taken into account.

Remarkably little literature exists on including information on the nominal scale nor on non-stationary interpolation methods. Journel (1983) developed Indicator Co-Kriging as a means to incorporate other data than the primary variable for spatial estimation. However, there needs to be an order in the data. Stein (1994) presented an early attempt to categorize possibilities of Co-Kriging. In a hydrogeological context, an early idea to include secondary information was presented by Ahmed et al. (1988), who developed External Drift Kriging (EDK). Monestiez et al. (1999) presented an attempt to include categorical variables as external drift. Other options include Co-Kriging (Matheron, 1971), or Simple Updating (SU, Bárdossy et al., 1996). All three methods lead to sharp edges at the interface between different types of secondary information. Additionally, these methods rely on Gaussian assumptions and take only the distance from a measurement into account, not the distribution of the conditioning measurements. Rivest and Marcotte (2012) interpolated groundwater solute concentrations by using flow-coordinates and non-stationary covariance functions. Brus et al. (2008) and Wibrin et al. (2006) used Maximum Entropy to predict soil categories in a computationally intensive despite low-dimensional approach. Liu et al. (2006) used categorical information from soil maps for spatial interpolation using Kriging. Emery and Silva (2009) used truncated Gaussian models to simulate cross-correlated fields between continuous and categorical variables in a mining application. They investigated the issue of hard or soft rock-type boundaries. In the case of soft boundaries (comparable to the influence of a certain neighborhood size in our case) they used cross-correlation, otherwise they simulated the grades separately.

This paper introduces two alternative methodologies, both based on multivariate spatial copulas. Results are compared with Gaussian copulas (B, Bárdossy, 2006), Ordinary Kriging (OK), and External Drift Kriging (EDK). If the composition of the land use in a neighborhood is considered, the size of the neighborhood is expressed in units of number of neighbors (*nneib*) of raster cells containing land use information in radial distance. Radii for circular shaped neighborhoods used commonly in this paper include ~ 1 km ($nneib = 32$), ~ 1.5 km ($nneib = 48$), ~ 2 km ($nneib = 64$), ~ 4 km ($nneib = 128$).

Data used and symbology are identical as in the related paper by Heisserer et al. (2016).

2. Spatial interpolation methods

This paper presents two novel interpolation methods based on multivariate copulas that are capable of including secondary information ("C1" and "C2", Table 1). Both alternatives are capable of describing spatial dependence and local influences, both methods are similar and differ only slightly in the underlying hypotheses. The first option (C1) is a fundamentally spatial process that is locally modified with prior information about the land use. The second option (C2) assumes that the quantiles of the marginal distribution are a continuous process. The marginal distribution can be evaluated at every location with an individual marginal distribution. Method C1 is the statistically more complete option and returns slightly superior cross-validation results. The methods are demonstrated for the estimation of groundwater quality parameters of anthropogenic origin, where the predominantly vertical infiltration processes are represented via the marginal distribution and the predominantly horizontal transport processes via the spatial dependence models.

The additional following methods for spatial estimation are assessed in this paper, for reference: A2, B, OK, and EDK. A schematic overview over the interpolation methods compared in this paper is given in Table 1. Method A2 does not describe spatial dependence; rather it maps the expected concentrations for locally mixed distributions that depend solely on the composition of the land use in the vicinity (Heisserer et al., 2016). Method B includes spatial dependence using copulas, but no secondary information. Additionally, the classical interpolation methods OK (without secondary information) and EDK (with secondary information) are assessed.

2.1. Without secondary information

In this paper, two methods are assessed that take a model of spatial dependence into account, but do not take advantage of secondary information. The reference method for geostatistical (as opposed to deterministic) estimation is OK that provides a least-squares estimate of the expected value and the expected variance at an arbitrary location, given a theoretical semivariogram model fitted to data.

Spatial estimation using copulas (method "B", Bárdossy and Li, 2008) maps the expected value of a conditional probability density that is based on both the distribution of measurement values of the conditioning points and the geometry of the observation network (Haslauer et al., 2008). In the basic case of interpolation method B, the measurement values are transformed into uniform space

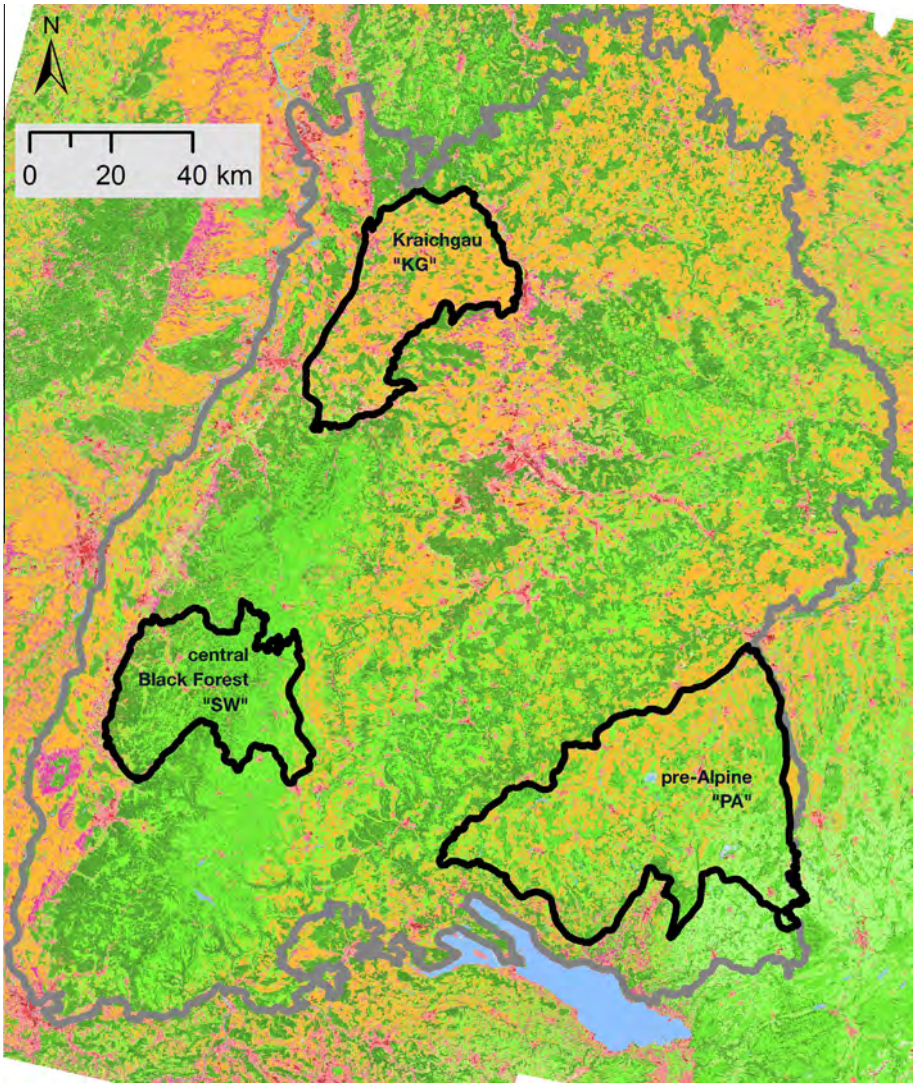

Fig. 1. Map of the interpolation domain, the state of Baden-Württemberg in south–west Germany (gray outline), and three selected areas (black outlines). The colors indicate land use categories (green colors indicate natural land use, orange represents agriculture, and red represents urban areas).

Table 1
Overview of interpolation methods used in this paper.

Method	Characteristics
	Concentration distributions based on...
A2	Composition of secondary information in the neighborhood; no spatial model; see Heisserer et al. (2016)
B	Spatial dependence and distributions of measurements; no secondary information
C1	Combination of "A2 and B" via prior information
C2	Combination of "A2 and B" via assumption of continuous marginals
	Reference methods; concentration mean and variance based on...
OK	Second order spatial dependence of measurements
EDK	Second order spatial dependence of measurements and secondary information

via the global nitrate distribution constructed by taking all measurements within the domain into account, independent of the dominant land use group near the measurement location. The distributions for each land use group deviate more and more from the global distribution, as a larger and larger neighborhood size is taken into account (Fig. 2). Censored measurements are taken into account via the probability of non-exceedance, as demonstrated by Bárdossy (2011).

2.2. Taking secondary information into account

This paper presents three methods that can take secondary information into account. All three methods require a rational number that represents the secondary information at all locations where a measurement exists, and at every location where a value is to be estimated. In this paper, such values represent the composition of the land use within a certain neighborhood. Here, they are the expected concentrations of locally mixed distributions (method A2, see Heisserer et al., 2016). Other sources for secondary information can be incorporated equally well: As an example, MODIS land cover is available globally and openly (Channan et al., 2014). Classification algorithms do exist, e.g. Hagner and Reese (2007) or Strahler (1980) up to a resolution of tens of meters on a regional scale.

The novel methods developed in this paper are copula-based methods C1 and C2. Both methods need the information necessary to construct the conditional probabilities for estimation to be available in uniform probability space. The two presented methods differ in how the information is transformed into uniform space before interpolation and transformed back into measurement space after interpolation. The locally mixed distributions (method

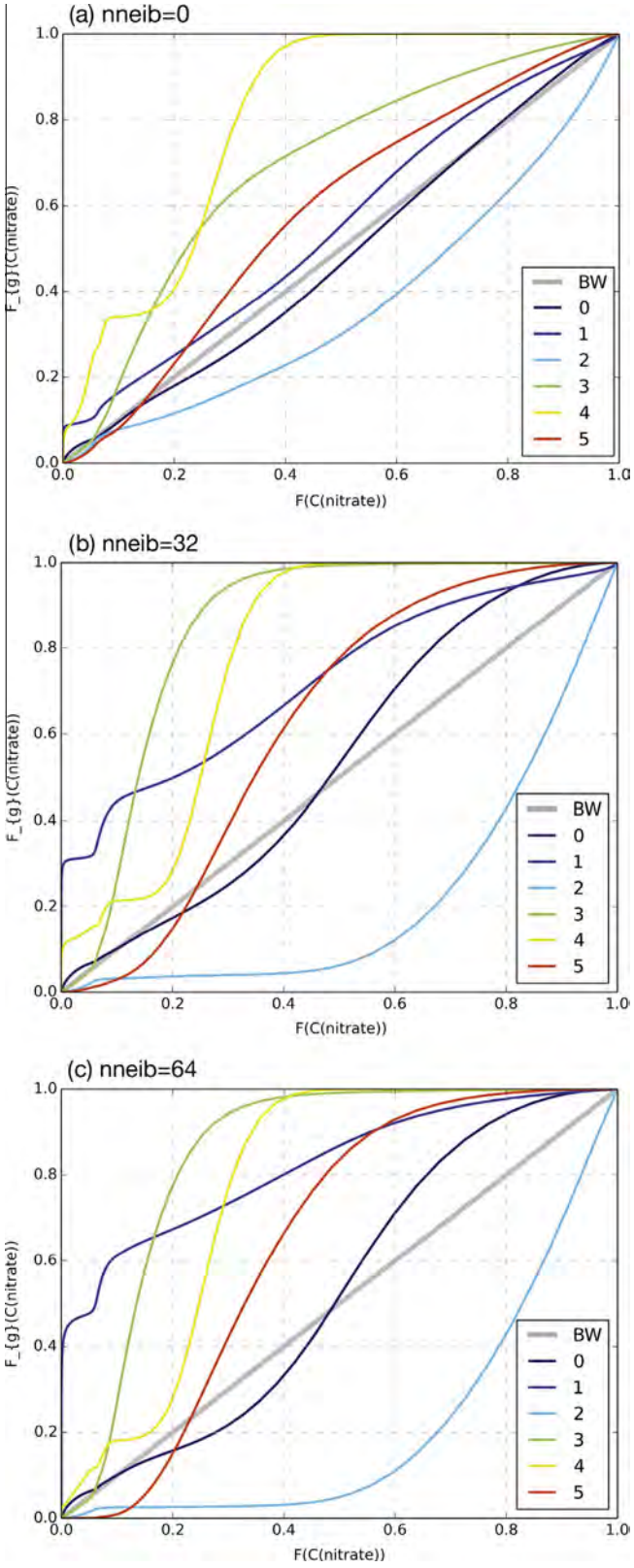

Fig. 2. Nitrate distributions in probability space for different neighborhood sizes. Each colored line indicates the distribution for a land use group. The gray line indicates the global nitrate distribution for the entire domain. The land use groups are identical to the ones presented by Heisserer et al. (2016). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A2) can be used to transform concentrations into uniform space (Fig. 2). Starting from zero neighborhood size to larger neighborhood sizes, the distributions start to diverge, as more information from a larger neighborhood size becomes available. For larger neighborhoods, the process of divergence is reversed.

As reference, the performance of EDK (Chilès and Delfiner, 2012) is assessed. The methodologies of OK and EDK are similar: in EDK, the Kriging matrix that needs to be solved for the least-square estimate contains one more row and column than in OK. This additional information at the observations and the right hand side is one dimension larger representing the secondary information at the interpolation location. The original concept of EDK takes no neighborhood into account, hence it corresponds to taking the secondary information via method A2 with $nneib = 0$ ("EDK 0") into account. This original concept is extended in this paper by using larger neighborhood sizes for EDK with method A2 as secondary information (e.g., "EDK 48").

2.2.1. Method C1

Method C1 models a spatial process with locally modified prior information. It uses information of the methods B and A2 at each interpolation location. The measurements are transformed into probability space for copula-based interpolation via the global distribution (as is done in method B). After the conditional probability is constructed, the back-transformation into measurement space is scaled via the law of total probability, taking information from method A into account (the locally mixed distributions reflecting the composition of land use in the neighborhood). The probability of a concentration to occur at location x_0 conditioned on the neighboring measurements in copula space and given the land use l_0 at the interpolation location is given in Eq. (1); details are presented in Appendix A.

$$P(U(x_0) = u_0 | U(x_k) = u_k, k = 1, \ldots, K, L(x_0) = l_0) = \frac{p_{t_0,l_0} P(U(x_0) = u_0 | U(x_k) = u_k, k = 1, \ldots, K)}{\sum_{r=1}^{K} p_{t_r,l_0} P\left(U(x_0) = \frac{r-\frac{1}{2}}{K} \middle| U(x_k) = u_k, k = 1, \ldots, K\right)} \quad (1)$$

2.2.2. Method C2

Method C2 is similar to method B; technically, the difference lies in the type of distribution functions both methods use for the transformations into and out of uniform probability space necessary for interpolation using copulas. In method C2, at every location, an individual marginal distribution is used that is based on the composition of the neighborhood (method A2), whereas for method B, the same distribution is used at every location (the global distribution of nitrate for Baden-Württemberg). The differences between methods B and C2 can be significant, up to $p \sim 0.5$: Fig. 2 shows that a distribution function value of $F_g(C) = 0.1$ can correspond to $F_{global,BW} \sim 0.1$ and $F_{local} \sim 0.6$ for the locally mixed distribution of group 2.

What this implies is that method C2 assumes that the quantiles of the marginal distribution are continuous. This is a strong assumption and might not be valid for small neighborhood sizes. Method C1 does not rely on this assumption.

3. Parameter estimation

The model parameters are estimated using a maximum likelihood-based approach described by Bárdossy and Li (2008). This estimation method results in reliable parameters for the covariance matrix (Fig. 3) that are discussed in this section. The estimated covariance parameters differ slightly whether censored

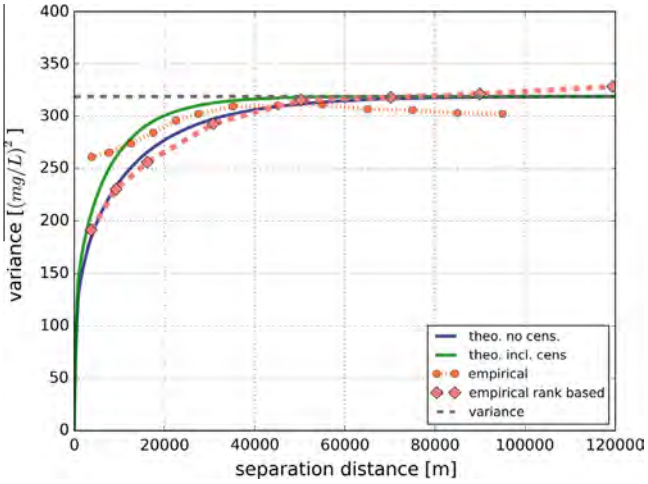

Fig. 3. Empirical and theoretical variograms for nitrate. The theoretical variogram that does not take censored measurements into account ("theo. no cens.", solid blue line) was used for OK and EDK. The theoretical variogram that does take censored measurements into account ("theo. incl. cens.", solid green line) was used for B, C1, and C2. The typical length scale is ~60 km. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

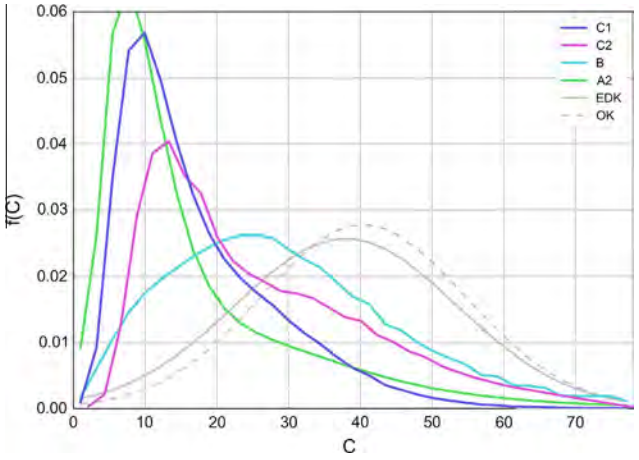

Fig. 4. An example of the interpolated densities of the presented methods at a randomly selected location within Baden-Württemberg. Methods A2, C1, and C2 were evaluated using a neighborhood size of *nneib* = 48, methods B, OK, and EDK use no neighborhood.

measurements are taken into account (as described by Bárdossy, 2011) or not. The difference is expected to be larger for variables with a larger proportion of censored measurements. Rank-based empirical variograms resolve the spatial structure better than measurement-based semi-variograms.

Solely based on bivariate statistics of the measurements, the structure of the semivariogram is difficult to estimate, particularly for short separation distances (orange dotted line on Fig. 3). The lack of a detectable structure of the semivariogram is likely attributable to the skewness in the data (1.55 for nitrate, 6.81 for chloride, see Heisserer et al., 2016). Rank-based semi-variograms are able to recover some parts of the semivariogram structure.

The estimated parameters for the dominantly horizontal process, as modeled via the spatial dependence structure, are significantly different compared to the typical scales of the dominantly vertical process of infiltration that is modeled via the marginal distributions. The typical length scale ("range") of the spatial dependence structure is ~ 40 km (Fig. 3). The ideal neighborhood size for mixed local distributions using method A2 is at *nneib* = 48, corresponding to a radius of ~ 1.5 km.

4. Results

The key outcome of spatial estimation is a distribution of estimated values at a given location. Fig. 4 shows the densities of nitrate concentrations for the estimation methods covered in this paper at one location in Baden-Württemberg. The strong impact of land use is clearly visible: the estimated densities based on methods that take land use into account differ strongly from the group of distributions that do not take land use into account.

The result of Kriging is an estimated mean and an estimated variance, which suffice to describe a Gaussian density function. With Copula-based approaches, the result is a full density function at any location. Based on this estimated function at every location an estimated value can be inferred. Typically that is the expected value. Other measures derived based on this distribution function are also possible, particularly for uncertainty assessment. Method A2 depends strongly on the composition of the neighborhood, but ignores spatial dependence, hence the results of A2 and Kriging based methods differ strongly. Interpolation method B is similar to Kriging as it does not consider secondary information; at the same time, it is different from Kriging as it considers the distribution of the measurement values and not only the geometry of the observation network. The methods proposed in this paper, C1 and C2 (that take both secondary information and spatial dependence into account) lie somewhere between methods A2 on one hand and Kriging based methods on the other hand. Besides from being symmetric, the output of OK and EDK do not vary much. It will be seen later that Kriging is the smoothest of the selected interpolators.

The interpolation methods are evaluated in detail using cross-validation (Section 4.1) and results are analyzed with the help of interpolated maps of the expected value, the standard deviation (a measure of uncertainty), and probability of exceedance of a threshold-concentration (Section 4.2).

4.1. Cross-validation

The proposed methods C1 and C2 perform best (Table 2) in leave-one-out cross-validation (Hawkins et al., 2003). The gains in performance compared to traditional estimation methods were achieved both in the estimation of the mean and in quantifying the uncertainty. The gains in the estimation of the mean are

Table 2
Overview of cross-validation results for interpolation methods B, C1, C2. For comparison, methods A2, OK, and EDK are shown. This table is for the contaminant nitrate. The measure "Δ80%" evaluates the performance of the model to express uncertainty: $\Delta 80\% = 80\% - p$ where p is the frequency of measured values being within the interpolated 80% quantile.

Method	*nneib*	MAE (C)	RMSE (C)	LEPS (–)	Δ80%
OK	0	8.7	13.2	0.153	−4.1
EDK	0	9.0	13.8	0.159	−3.9
EDK	48	8.6	13.2	0.151	−4.7
A2	0	12.1	16.3	0.231	−7.2
A2	64	10.4	14.7	0.192	−5.1
B	-	9.1	13.3	0.157	−2.7
C1	0	8.9	13.0	0.151	−2.2
C2	0	9.6	14.0	0.163	−1.3
C1	32	8.3	12.3	0.139	−1.7
C2	32	8.6	13.0	0.145	−1.2
C1	48	8.2	12.3	0.136	−1.8
C2	48	8.5	13.0	0.142	−1.2
C1	64	8.2	12.3	0.136	−1.9
C2	64	8.5	13.0	0.142	−2.0
C1	128	8.2	12.4	0.137	−2.7
C2	128	8.6	13.3	0.142	−2.7

recognizable in concentration space (MAE, RMSE) and in probability space (LEPS). The definition of the MAE, RMSE, and LEPS is given in Eqs. (2)–(4), respectively.

$$\mathrm{MAE} = 1/n \cdot \sum_{i=1}^{n} (|c_{mess}(i) - c_{interpol}(i)|) \tag{2}$$

$$\mathrm{RMSE} = \sqrt{1/n \cdot \sum_{i=1}^{n} (c_{mess}(i) - c_{interpol}(i))^2} \tag{3}$$

$$\mathrm{LEPS} = 1/n \cdot \sum_{i=1}^{n} (F[c_{mess}(i)] - F[c_{interpol}(i)]) \tag{4}$$

A decrease of the RMSE from 13.2 using OK, to 12.3 using method C1 with at least some neighborhood size, explains $\sim 6\%$ of the variability inherent in the OK estimations. It has been shown that method A2 with $nneib = 48$ reduces the variability compared to using the global mean (method A1) by $\sim 25\%$ (Heisserer et al., 2016). OK reduces this variability by another $\sim 25\%$ and method C1 reduces the variability by the above-mentioned $\sim 6\%$, using also $nneib = 48$. Starting from the global mean, using the best performing method C1, a reduction in variability of estimation of $\sim 50\%$ could be achieved. This reduction is similar for MAE, and even higher at $\sim 70\%$ for the LEPS. Methods C1 and C2 perform about equally well, however method C1 performs consistently slightly better than C2.

Copula-based interpolation methods perform particularly well in quantifying the uncertainty. To evaluate the quality of the uncertainty measure, the percentage of interpolated values that fall into the 80% confidence interval was evaluated. In a perfect system, 80% of the interpolated values would fall into this 80% confidence interval. Using this measure of deviation from the 80% confidence interval, a reduction of $\sim 3.5\%$ in uncertainty could be achieved by switching from Kriging based methods to copula-based methods that take secondary information into account (either method C1 or C2 and a neighborhood size of $nneib = 48$).

Detailed cross-validation results are listed in Table 2 and are summarized here. Leave-multiple-out or leave-clusters-out did not give significantly differing results. As soon as we take some information about the composition of the neighborhood into account ($nneib > 0$), cross-validation results improve significantly. The added benefit of larger neighborhood sizes is comparatively small. For large neighborhood sizes, there is a similar cap: for example a neighborhood size of $nneib = 128$ performs worse than a neighborhood size of $nneib = 64$. This means that there is an optimum neighborhood size and that the information contained in the neighborhood approaches the information contained in the entire domain, as neighborhood sizes increase. The cross-validation results for estimating the mean seem to point to an optimum between $nneib = 48$ and $nneib = 64$. The cross-validation results for quantifying the uncertainty seem to point to an optimum of slightly smaller neighborhood sizes of $\sim nneib = 32$.

The most important outcome of this analysis is to take some secondary information at all into account: The biggest increase in performance was achieved with method A2 when considering at least some neighborhood, when moving from a neighborhood size of zero to a size of $nneib = 64$. Despite the fact that method A2 does not take spatial dependence into account it performs surprisingly well. Although RMSE explains $\sim 6\%$ of the variability, it is not an ideal measure for performance, as $\sim 80\%$ of the RMSE value is contributed by only the upper $\sim 20\%$ of the estimates.

In the traditional sense of EDK, when secondary information is taken into account only directly at the measurement locations, then EDK performs worse than OK. This means that taking the land use information into account via EDK is not a good strategy. Additionally, in that case, 15 values were interpolated as unphysical negative concentrations. When using the expected concentration based on the composition of a neighborhood via method A2 ("EDK 48" stands for EDK with secondary information obtained from A2 with a neighborhood size of $nneib = 48$), the performance of EDK improves. In the case of EDK 48, all three measures (MAE, RMSE, and LEPS) improve and the result is slightly better than with OK. Method B performs similar to OK in estimating the mean, but much better in quantifying the uncertainty.

In the direct comparison of estimated cross-validation results with the measured concentrations of all analyzed interpolation methods, there is a clear trend towards best performance of the most sophisticated interpolation methods. Method C1 performs best with a correlation coefficient of $r = 0.75$, slightly better than method C2 ($r = 0.73$). Method A2 that does not include a geostatistical model performs worst with a correlation coefficient of $r = 0.54$. Methods with geostatistical model but without secondary information perform in the middle. The typical problem of spatial estimation persists, large values are underestimated, small values are overestimated.

This comparison between interpolation methods was investigated further by looking point-wise at all pairs of the presented interpolation methods and counting which one performs better than the other. Based on those frequencies, probabilities of a method winning against every other method can be inferred. It is striking how well method C1 performs: it wins more than half of the time against every other interpolation method (Fig. 5). Interestingly, the best competitor against C1 is not C2 but method A2. Also, method A wins against any Kriging based method.

4.2. Spatial estimation

The cross-validation results are important, but looking at the spatial distribution of the performance of spatial estimation methods provides another important aspect. A reduction of 3.5% in deviation from the 80% confidence interval is an important feature, however, looking at the structure of the uncertainty provides more insight into the performance of an interpolation method. The typical result of spatial estimation are maps, particularly maps of the expected value of the estimated variable. Complementary information about the quality of the estimation procedure can be obtained by mapping the standard deviation of the estimated variable and the probability of exceedance of a threshold. Fig. 6 shows these three types of maps for nitrate concentration within Baden-Württemberg (represented in the columns) for the

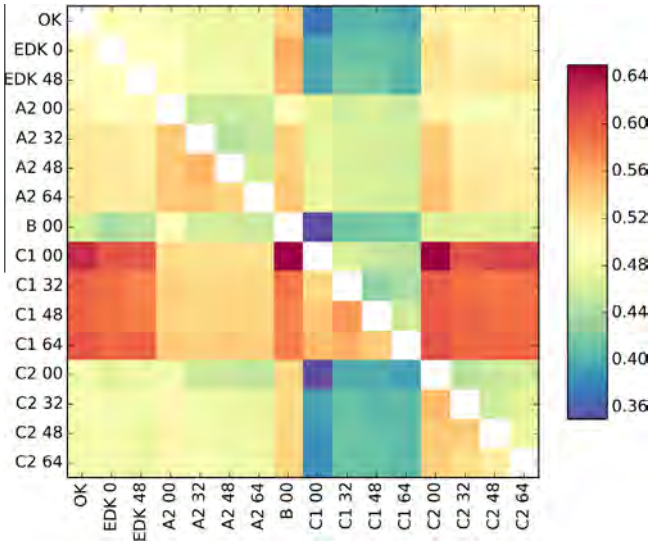

Fig. 5. Probabilities of winning [–] when comparing one interpolation method against every other interpolation method pairwise.

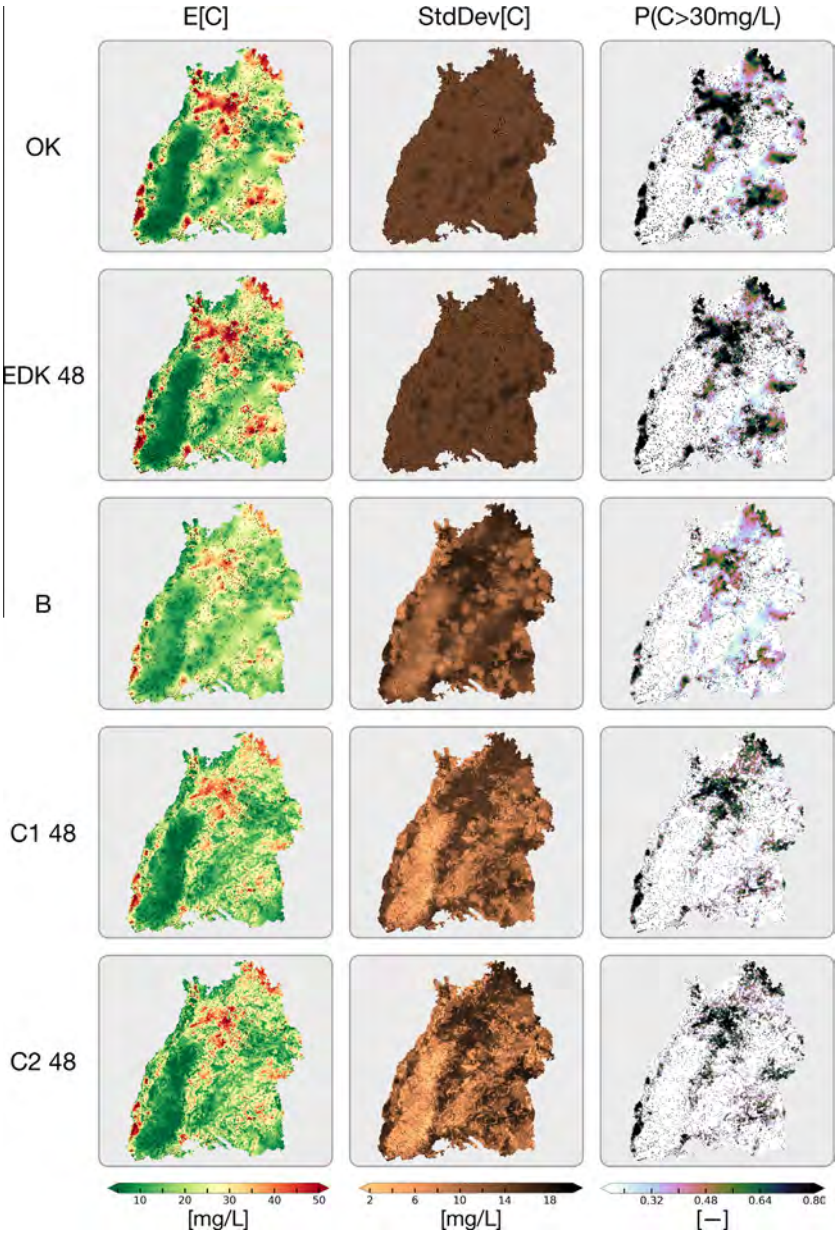

Fig. 6. Interpolated maps for nitrate in Baden-Württemberg. The three different types of maps are arranged in columns (expected value *E*, standard deviation of the estimate StDev, and probability of exceedance of a nitrate concentration of $C = 30$ mg/L). Different interpolation methods are arranged in rows: OK, EDK, B, C1 and C2. EDK, C1, and C2 are calculated based on a neighborhood size of $nneib = 48$.

presented interpolation methods (represented in the rows). This figure is available in higher resolution in Supplemental material.

Interpolation methods C1 and C2 lead to the most heterogeneous maps. This can be expected as they take both the land use information and horizontal spatial dependence into account. The Black Forest in the south-west of the domain (the dominantly green area on Fig. 1, extending the area of the central Black Forest northward and southward) is an interesting area to consider: the measurement values of nitrate under the Black Forest are homogeneously small, as the Black Forest is a large area with forest being the dominant land use. This is the reason why OK and EDK (which only looks at the land use directly at the measurement location) estimate a continuous area (without holes) with homogeneously low nitrate concentrations. Despite the generally homogeneous forest, valleys with agriculture and towns also exist. In these valleys one would expect higher nitrate concentrations than under the larger forested areas, even if only a few measurements existed in the valleys. The methods C1 and C2 reflect this small-scale heterogeneity in the land use in the estimated nitrate concentrations.

The methods that take not only the geometry of the observation network into account, but also the measurement values (B and C), have a much improved measure of uncertainty. As for the other methods presented in this paper, they also lead to small uncertainties in the vicinity of measurement locations. Additionally, methods C1 and C2 take the land use composition into account. This leads to further decreased uncertainty. Areas with both homogeneous measurements and homogeneous land use (such as the Black Forest) lead to the smallest uncertainty. However, in smaller-scale features with heterogeneous land use, such as the valleys in the Black Forest, uncertainty is large. These properties make methods C1 and C2 the most plausible interpolation methods. The Kriging-based uncertainty estimates are too low and unrealistic, because the structure of the error is lost completely. The

Table 3
Overview of descriptive statistical measures of the interpolated expected values for selected areas within the domain and for selected interpolation methods. "ID" indicates the name of the sub-domain of the interpolation (the full domain of Baden-Württemberg (BW), a pre-Alpine area (PA), the central Black Forest (SW), and the Kraichgau (KG), see also Fig. 1).

ID	Method	*nneib*	Mean	stdD	Skew
BW	OK	–	20.9	11.6	0.80
BW	EDK	0	21.0	12.8	0.96
BW	EDK	48	20.9	12.5	0.96
BW	A2	48	19.3	7.7	0.26
BW	B	48	19.5	7.2	0.68
BW	C1	48	19.1	9.7	0.61
BW	C2	48	19.4	10.2	0.68
PA	OK	–	26.7	8.7	0.39
PA	EDK	0	27.0	10.4	0.39
PA	EDK	48	26.7	9.4	0.41
PA	A2	48	19.5	6.8	0.05
PA	B	48	22.2	5.9	0.48
PA	C1	48	21.3	7.7	0.32
PA	C2	48	21.4	8.5	0.35
KG	OK	–	32.8	11.0	0.25
KG	EDK	0	32.8	11.5	0.28
KG	EDK	48	33.0	13.0	0.70
KG	A2	48	27.1	5.8	−0.49
KG	B	48	26.9	6.9	0.09
KG	C1	48	30.4	8.4	−0.17
KG	C2	48	29.6	9.2	0.15
SW	OK	–	8.3	4.0	1.77
SW	EDK	0	8.2	4.1	2.02
SW	EDK	48	8.0	3.8	1.59
SW	A2	48	11.6	4.1	1.18
SW	B	48	12.7	2.6	0.75
SW	C1	48	9.0	3.5	1.20
SW	C2	48	9.6	3.7	1.08

uncertainty of the interpolated estimates using OK and EDK is a reflection of the geometry of the observation network. The largest standard deviation occurs in locations that are the farthest away from a measurement location. Even if only slightly, the uncertainty of EDK is larger than of OK, despite the fact that EDK uses secondary information. Methods C1 and C2 do not only lead to superior cross-validation results but also to an improved spatial structure of both the estimated mean and the estimate of uncertainty of the interpolation methods.

The probability that a threshold concentration is exceeded is important information for people that need to make decisions based on interpolated concentration maps. The probabilities of exceeding a concentration of 30 mg/L are shown on the right column of Fig. 6. The patterns are similar to the ones for the expected values: Methods C1 and C2 show the largest heterogeneity; the deviations between different interpolation methods are largest where the effects of land use are strongest.

The arithmetic mean of the estimated expected nitrate concentrations varies by 1.9 mg/L between the presented interpolation methods when taking Baden-Württemberg as the interpolation domain. With $\sim 45,000$ estimates, this difference is significant. Due to the best performance of method C1 in cross-validation and its superior spatial structure, both for estimated values and uncertainty estimates, it can be concluded that the mean of 19.1 mg/L is the best estimate (Table 3). As methods C1 and C2 include land use information and spatial dependence, the standard deviation over all estimates in the domain is slightly increased compared to methods B or A2. Generally, the order of the mean of the estimates is identical to the order of the performance of the estimation methods. This is certainly true for the entire domain of Baden-Württemberg, but also for other sub-domains (Fig. 1). However, things are different in the central Black Forest, where methods C1 and C2 suggest a larger mean than OK. Likely, the reason is again that those two methods take both land use and spatial dependence into account. The increased mean in the central Black Forest might hence be attributable to the agricultural areas in the large valleys.

With the maps analyzed so far it is not possible to highlight differences between the presented interpolation methods. The following two figures show differences between the interpolation methods in expected concentration (Fig. 7) and expected standard deviation of concentration (Fig. 8). In these maps of differences, the different hypotheses dominate. It becomes clear that OK and EDK are similar, C1 and C2 are similar, and both A and B appear unique.

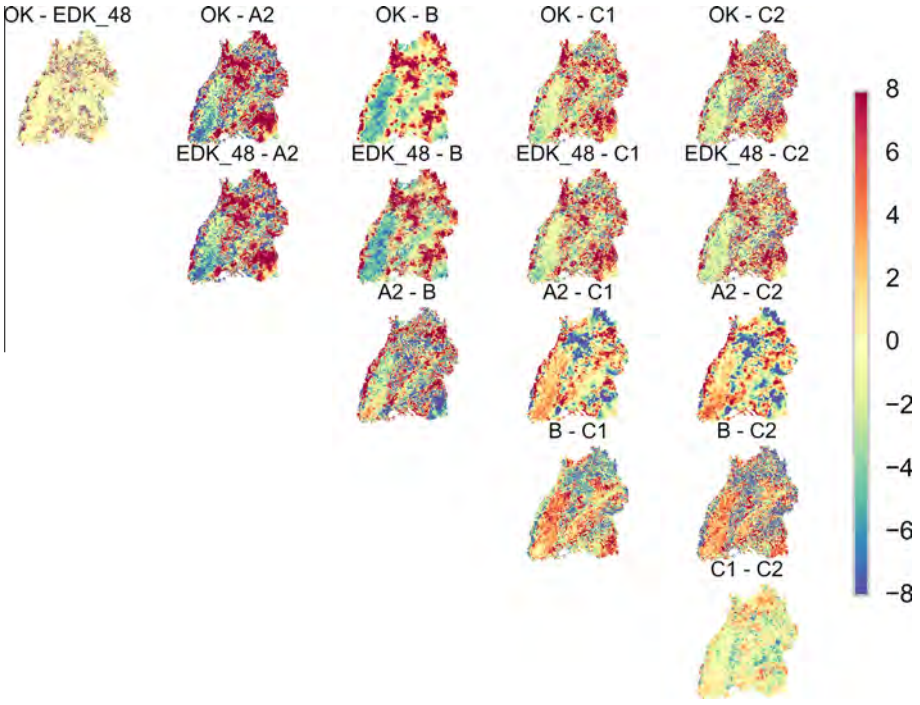

Fig. 7. Differences in estimated nitrate concentrations [mg/L] between selected interpolation methods.

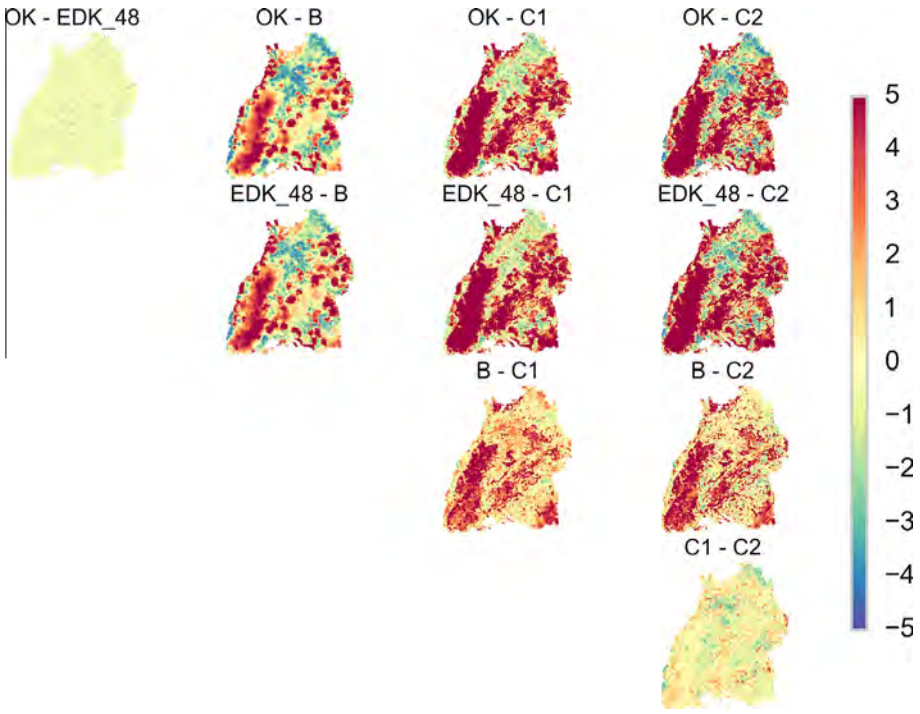

Fig. 8. Differences in standard deviations [mg/L] of nitrate concentrations between selected interpolation methods.

The patterns in the map of differences between exceedance probabilities is similar to the map in differences between expected standard deviations (not shown).

The differences in expected concentrations are largest in areas where measurements are either particularly homogeneous or particularly heterogeneous, for example in the Black Forest (as noted in the previous section). The difference between the Kriging methods OK and EDK and the other methods presented are relatively large: Kriging takes the geometric configuration of the observation network into account. The largest differences occur between Kriging which considers only the measurement locations, and method A2, which considers only the land use, not any spatial dependence, and is hence spatially distributed differently. Accordingly, the differences between method B and methods C1 and C2 are smaller than between Kriging and methods C1 and C2. Method B incorporates the degree of homogeneity of the measurement field by including the measurement values. This figure showing this improvement is important, because method B did not perform significantly better than Kriging in cross-validation (Table 2). The pairs (OK, EDK) and (C1, C2) are similar. Each pair have very similar underlying hypotheses. This similarity is evident also in the maps of differences in estimation standard deviation and in the differences in probabilities of exceeding a concentration of 30 mg/L.

The difference between Kriging-based estimation methods and methods C1 and C2 in estimating the probability of exceedance of a concentration of $C = 30$ mg/L is larger than 20% over the domain of Baden-Württemberg. The important thing to remember is that those differences are not the same everywhere, and neither scaled everywhere in the same way, but are a reflection of the conditions in specific areas. This can be considered to be significant advancement.

5. Conclusions

Locally spatially averaged information can be used effectively in a copula-based spatial estimation approach and performs significantly better than other Kriging based approaches (OK, EDK). The presented method uses secondary information to construct non-stationary maps and includes censored measurements in a full stochastic framework. Using this framework leads to more "realistic" maps, meaning here that the interpolated values exhibit a higher degree of heterogeneity compared to conventional methods, as well as smooth transitions between areas of different secondary information. The latter fact can be attributed to the incorporation of land used based on a neighborhood with a certain size. With the given data set of nitrate concentrations in Baden-Württemberg best results were obtained with method C1 and a neighborhood size of $nneib = 48$ (circular neighborhood with radius ~ 1.5 km).

The superiority of copula-based methods goes beyond cross-validation results: uncertainty estimates, quality of the uncertainty, and the spatial structure of the uncertainty are much improved to traditional geostatistical approaches. When evaluating the uncertainty of the interpolation, copula-based methods win in most of the cases. This is a clear indication that the spatial model of the copula models is improved compared to traditional methods. This type of performance is not only visible in cross-validation results, but also in the spatial structure of estimated concentrations. The heterogeneity in the secondary information is reflected in the interpolation as well as the structure of the measurements: Large continuous areas of homogeneous measurements together with homogeneous land use result in the smallest uncertainty of spatial estimation. With the proposed methods, the structure of the estimation is not only dependent on the geometry of the observation network, but also on the full distribution of the measured values, and hence the gradient of the measured values. This property becomes evident as larger homogenous areas are less uncertain compared to smaller isolated patches.

A general problem of spatial estimation is the bias that is inherent in most measurement networks: measurements are generally not taken at locations in correspondence to the proportion of the occurrence of land use groups. The average of the measured nitrate concentrations within Baden-Württemberg is 19.9 mg/L. If the

measurements are corrected for the frequency of land uses occurring within the interpolation domain, the average increases to 21.1 mg/L. If the measurements are corrected for the frequency of land uses occurring directly at the measurement locations, the average increases to 21.2 mg/L. This difference is significant based on $\sim 45,000$ estimates. Hence, it is necessary to include the information of land use into the estimation of groundwater quality parameters. In this paper, we have presented methods that can accommodate such differences. A difference of ~ 2 mg/L in estimated average nitrate concentration over the domain of Baden-Württemberg or a difference of $\sim 20\%$ in probability of exceeding a threshold of $C = 30$ mg/L is a significant conclusion! Furthermore, we have shown that taking those biases into account in a reasonable way is prerogative on smaller sub-domains. Then, these differences between interpolation methods amplify and taking land use into account becomes mandatory.

A lot of effort has been spent to decrease the RMSE by about one concentration unit and the variability of estimation by about 10%. The spatial structure of the estimated concentrations and uncertainties is also much more reasonable. One might argue: "nice, but so what?" In a bootstrap procedure, sequentially more and more measurements were left out in the leave-one-out cross-validation procedure. Accordingly, the measures of MAE, RMSE, and LEPS increase. The artificial and random removal of 500 measurements lead to an increase of the RMSE of maximally ~ 0.25 mg/L when OK was used. On the other hand, method C1 reduced the RMSE from OK to C1 with optimal neighborhood size by ~ 1.0 mg/L. Based on this rough estimation, certainly a few hundred additional measurements, likely more than 10% additional measurements compared to existing number of measurements, would be necessary when the same performance in terms of RMSE should be achieved with OK as is achieved with method C1 – at a significant larger cost for installation, maintenance, and sample analysis.

The presented method is flexible and other or additional types of secondary information could be included. The presented method would work with smaller data-sets, however adjustments to the classification would be necessary. The presented method is able to estimate the impacts of land use on groundwater quality at a snapshot in time. It could be repeated for multiple snapshots in time, and could offer then a tool to quantify the effects of land use changes on groundwater quality.

Acknowledgements

Thanks to the state office for the environment, measurement, and natural protection of the state of Baden-Württemberg (LUBW), Germany, for collaboration. This work was funded by German Research Foundation (DFG) Grant GRK 1829/1 and DFG Grant HA 7339/2-1.

Appendix A. Derivation of conditional probability for interpolation-method C1

The goal is to find the conditional distribution of $U(x_0)$ given the observations $U(x_k) = u_k$ for $k = 1, \ldots, K$, and the land use at point x_0 (Eq. (A.1)). The derivation of this approach is demonstrated in the following.

Assume, the variable U is uniformly distributed (discrete case, Eq. (A.2)). Incorporating the additional information of the land use at location l_k leads to a modified conditional distribution (Eq. (A.3)). The approach for this local distribution is explained by Heisserer et al. (2016).

The goal is to calculate the probability given in Eq. (A.1).

$$P(U(x_0) = u | U(x_k) = u_k, \quad k = 1, \ldots, K, \quad L(x_0) = l_0) \tag{A.1}$$

$$P\left(U(x_k) = \frac{t - \frac{1}{2}}{K}\right) = \frac{1}{K} \quad k = 1, \ldots, K, \quad t = 1, \ldots, K \tag{A.2}$$

$$P\left(U(x_k) = \frac{t - \frac{1}{2}}{K} | L(x_k) = g\right) = p_{t,g} \quad k = 1, \ldots, K \tag{A.3}$$

thus

$$\sum_{t=1}^{K} p_{t,g} = 1 \tag{A.4}$$

The spatial copula is:

$$P\left(U(x_0) = \frac{t - \frac{1}{2}}{K} \middle| U(x_k) = \frac{k - \frac{1}{2}}{K}\right) = q_{t,k,x_0,x_k} \tag{A.5}$$

thus

$$\sum_{t=1}^{K} q_{t,k,x_0,x_k} = 1 \tag{A.6}$$

Each land use class has its own frequency of occurrence

$$P(L(x_k) = g) = v_g \quad k = 1, \ldots, K \tag{A.7}$$

thus

$$\sum_{g=1}^{G} v_g = 1 \tag{A.8}$$

$$P(U(x_0) = u_0 | U(x_k) = u_k, k = 1, \ldots, K, L(x_0) = l_0) = \tag{A.9}$$

$$= \frac{P(U(x_0) = u_0, U(x_k) = u_k, k = 1, \ldots, K, L(x_0) = l_0)}{P(U(x_k) = u_k, k = 1, \ldots, K, L(x_0) = l_0)} \tag{A.10}$$

$$\begin{aligned} &= \{P(L(x_0) = l_0 | U(x_0) = u_0, U(x_k) = u_k, k = 1, \ldots, K) \\ &\quad \cdot P(U(x_0) = u_0, U(x_k) = u_k, k = 1, \ldots, K)\} \\ &\quad / P(L(x_0) = l_0 | U(x_k) = u_k, k = 1, \ldots, K) P(U(x_k) = u_k, k = 1, \ldots, K) \end{aligned} \tag{A.11}$$

$$\begin{aligned} &= \{P(L(x_0) = l_0 | U(x_0) = u_0, U(x_k) = u_k, k = 1, \ldots, K) \\ &\quad / P(L(x_0) = l_0 | U(x_k) = u_k, k = 1, \ldots, K)\} \\ &\quad \cdot P(U(x_0) = u_0 | U(x_k) = u_k, k = 1, \ldots, K) \end{aligned} \tag{A.12}$$

$$\begin{aligned} &= \frac{P(L(x_0) = l_0 | U(x_0) = u_0)}{P(L(x_0) = l_0 | U(x_k) = u_k, k = 1, \ldots, K)} P(U(x_0) = u_0 | U(x_k) \\ &= u_k, k = 1, \ldots, K) \end{aligned} \tag{A.13}$$

Looking at the terms one by one:

$$P(U(x_0) = u_0 | U(x_k) = u_k, k = 1, \ldots, K) \tag{A.14}$$

is the pure copula interpolation for x_0 without land use.

$$P(L(x_0) = l_0 | U(x_0) = u_0) = \tag{A.15}$$

$$P(U(x_0) = u_0 | L(x_0) = l_0) \frac{P(L(x_0) = l_0)}{P(U(x_0) = u_0)} = \tag{A.16}$$

$$p_{t_0,l_0} \frac{v_{l_0}}{\frac{1}{K}} \tag{A.17}$$

$$\begin{aligned} &P(L(x_0) = l_0 | U(x_k) = u_k, k = 1, \ldots, K) \\ &= \sum_{r=1}^{K} P\left(L(x_0) = l_0 | U(x_0) = \frac{r - \frac{1}{2}}{K}\right) P\left(U(x_0) = \frac{r - \frac{1}{2}}{K} \middle| U(x_k) = u_k, k = 1, \ldots, K\right) \\ &= \sum_{j_r=1}^{m} p_{t_r,g} \frac{v_g}{\frac{1}{K}} P\left(U(x_0) = \frac{r - \frac{1}{2}}{K} \middle| U(x_k) = u_k, k = 1, \ldots, K\right) \end{aligned} \tag{A.18}$$

Finally giving the probability:

$$\frac{p_{t_0,l_0}\frac{v_{l_0}}{\frac{1}{K}}P(U(x_0)=u_0|U(x_k)=u_k,k=1,\ldots,K)}{\sum_{r=1}^{K}p_{t_r,l_0}\frac{v_{l_0}}{\frac{1}{K}}P(U(x_0)=\frac{r-\frac{1}{2}}{K}|U(x_k)=u_k,k=1,\ldots,K)}$$
$$=\frac{p_{t_0,l_0}P(U(x_0)=u_0|U(x_k)=u_k,k=1,\ldots,K)}{\sum_{r=1}^{K}p_{t_r,l_0}P\left(U(x_0)=\frac{r-\frac{1}{2}}{K}\middle|U(x_k)=u_k,k=1,\ldots,K\right)} \quad (A.19)$$

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jhydrol.2016.01.054.

References

Ahmed, S., de Marsily, G., Talbot, A., 1988. Combined use of hydraulic and electrical properties of an aquifer in a geostatistical estimation of transmissivity. Ground Water 26, 78–86.

Bárdossy, A., 2006. Copula-based geostatistical models for groundwater quality parameters. Water Resour. Res. 42.

Bárdossy, A., 2011. Interpolation of groundwater quality parameters with some values below the detection limit. Hydrol. Earth Syst. Sci. 15, 2763–2775.

Bárdossy, A., Haberlandt, U., Grimm-Strehle, J., 1996. Interpolation of groundwater quality parameters using additional information. In: Soares, A., Gómez-Hernández, J.J., Froidevaux, R. (Eds.), geoENV I – Geostatistics for Environmental Applications; Proceedings of the Geostatistics for Environmental Applications Workshop. Kluwer Academic Publishers, Lisbon, Portugal, pp. 189–200.

Bárdossy, A., Li, J., 2008. Geostatistical interpolation using copulas. Water Resour. Res. 44.

Bárdossy, A., Pegram, G., 2013. Interpolation of precipitation under topographic influence at different time scales. Water Resour. Res. 49, 4545–4565.

Brus, D.J., Bogaert, P., Heuvelink, G.B.M., 2008. Bayesian Maximum Entropy prediction of soil categories using a traditional soil map as soft information. Eur. J. Soil Sci. 59, 166–177.

Channan, S., Collins, K., Emanuel, W.R., 2014. Global mosaics of the standard MODIS land cover type data. Electronic Source: <http://glcf.umd.edu/data/lc/>. Access date: 2016/01/07.

Chilès, J.P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty, second ed. Wiley.

Emery, X., Silva, D.A., 2009. Conditional co-simulation of continuous and categorical variables for geostatistical applications. Comput. Geosci. 35, 1234–1246.

Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. J. Hydrol. 228, 113–129.

Hagner, O., Reese, H., 2007. A method for calibrated maximum likelihood classification of forest types. Rem. Sens. Environ. 110, 438–444.

Haslauer, C.P., Li, J., Bárdossy, A., 2008. Application of copulas in geostatistics. In: Atkinson, P.M., Lloyd, C.D. (Eds.), geoENV VII – Geostatistics for Environmental Applications. Springer, Dordrecht, pp. 395–404.

Hawkins, D.M., Basak, S.C., Mills, D., 2003. Assessing model fit by cross-validation. J. Chem. Inform. Model. 43, 579–586.

Heisserer, T., Haslauer, C.P., Bárdossy, A., 2016. Including land-use information for the spatial estimation of groundwater quality parameters – 1. Local estimation based on neighbourhood composition. J. Hydrol.

Journel, A.G., 1983. Non-parametric estimation of spatial distributions. Math. Geol. 15, 445–468.

Liu, T.L., Juang, K.W., Lee, D.Y., 2006. Interpolating soil properties using kriging combined with categorical information of soil maps. Soil Sci. Soc. Am. J. 70, 1200.

Matheron, G., 1971. The Theory of Regionalized Random Variables and its Applications. Centre de Morphologie Mathématique de Fontainebleau.

Monestiez, P., Allard, D., Sanchez, I.N., Courault, D., 1999. Kriging with categorical external drift: use of thematic maps in spatial prediction and application to local climate interpolation for agriculture. In: Gómez-Hernández, J.J., Soares, A., Froidevaux, R. (Eds.), Geostatistics for Environmental Applications, geoENV II. Springer, Netherlands, Dordrecht, pp. 163–174.

Pozdnyakova, L., Zhang, R., 1999. Geostatistical analyses of soil salinity in a large field. Prec. Agric. 1, 153–165.

Rivest, M., Marcotte, D., 2012. Kriging groundwater solute concentrations using flow coordinates and nonstationary covariance functions. J. Hydrol., 238–253

Shim, B., Chung, S., Kim, H., Sung, I., 2004. Intrinsic random function of order *k* kriging of electrical resistivity data for estimating the extent of saltwater intrusion in a coastal aquifer system. Environ. Geol. 46, 533–541.

Stein, A., 1994. The use of prior information in spatial statistics. Geoderma 62, 199–216.

Strahler, A.H., 1980. The use of prior probabilities in maximum likelihood classification of remotely sensed data. Rem. Sens. Environ. 10, 135–163.

Wibrin, M.A., Bogaert, P., Fasbender, D., 2006. Combining categorical and continuous spatial information within the Bayesian maximum entropy paradigm. Math. Geosci. 20, 423–433.

Zhu, H., Journel, A.G., 1993. Formatting and integrating soft data: stochastic imaging via the Markov–Bayes algorithm. In: Geostatistics Tróia 1992. Springer, Netherlands, Dordrecht, pp. 1–12.